CEB Section Experiment Archiving Tool

A Knowledge Management Approach to Designing and Archiving Experiments and Their Data

Marc S. Grushcow
NTT SYSTEMS INC.

PWGSC Contract No. W7711-0- 7687
Call-Up 17

On behalf of
DEPARTMENT OF NATIONAL DEFENCE

as represented by
Defence Research and Development Canada - Toronto
1133 Sheppard Avenue West
Toronto, Ontario, Canada
M3M 3B9

DRDC Scientific Authority
Andrea Hawton
(416) 635-2100 x 3036

Defence R&D Canada – Toronto

Contract Report

DRDC Toronto CR 2004-069

March 2004

Author

Marc S. Grushcow

Approved by

Andrea Hawton
Scientific Authority

Approved for release by

K. M. Sutton
Chair, Document Review and Library Committee

© Her Majesty the Queen as represented by the Minister of National Defence, 2004
© Sa majesté la reine, représentée par le ministre de la Défense nationale, 2004

Abstract

The Command Effectiveness and Behaviour Section at DRDC Toronto has been conducting cognitive psychology and team related experiments since the late 1970's. Over that time it has amassed a large volume of raw and analyzed data from a series of related, large and small research studies. Technically, data from different studies can be combined for analysis because they are all sufficiently well identified with respect to experiment, subject, task, trial, etc. However, this has not turned out to be the case. The reason is that, particularly for the early experiments, there is no longer sufficient documentation, or corporate knowledge, to reconstruct the experimental designs. This absence of knowledge of the context in which the data were collected renders them useless for the purpose of further analysis.

The reliance on staff as the repository of corporate knowledge is not unusual. But the consequences in research environments, particularly where public funding is involved, are far reaching and involve complex relationships between researchers, support staff, research organizations, and their funding agencies. The report begins by reviewing this situation and lays out a modest approach that protects corporate knowledge and experimental data while addressing the needs of each group. The central features of the approach are codification of experiment metadata and their representation in XML. XML is used since it is recognized as the standard mechanism for representing, transforming, and sharing computer-based descriptive information.

The approach is implemented in an application called *PI Assist*. The report presents a brief overview of the system and then discusses the results of its first field trial. This identified a number of enhancements that were implemented for the current version that is now ready for the second field trial.

In the course of pursuing this work, NTT also investigated approaches to searching collections Rof *PI Assist* documents. An implementation using the ORACLE database management system was designed and implemented. This was not altogether satisfactory, primarily for reasons relating to the relative immaturity of the combined database/XML technology. Since there is no current need for this search facility, owning to the small number of documents available, we recommend suspending work in this area until third party solutions evolve further.

The report concludes by identifying the need for further field trials with scientists and recommending that support staff in CEB begin using *PI Assist* to document experiments that have already been completed.

105-05737/

Résumé

La Section de l'efficacité du commandement et du comportement à RDDC Toronto mène depuis la fin des années 1970 des expériences liées aux équipes et à la psychologie cognitive. Pendant cette période, elle a accumulé un volume important de données brutes et analysées à l'aide d'une série d'études de recherche connexes, de grande et de faible envergures. Sur le plan technique, il est possible de combiner les données provenant d'études différentes, car l'expérience, l'objet, la tâche, les essais, etc. de celles-ci sont assez bien définies. Par contre, cela n'a pas été possible ici. En effet, en ce qui concerne plus particulièrement les premières expériences, il n'y a plus assez de documentation, ou de savoir collectif, pour reconstruire les concepts expérimentaux. Étant donné un tel manque de contexte quant à la façon dont les données ont été recueillies, il est donc inutile d'analyser davantage celles-ci.

Il n'est pas rare de recourir et se fier au personnel pour qu'il joue le rôle de dépôt du savoir collectif. Toutefois, les conséquences d'une telle décision dans des environnements de recherche, en particulier lorsque des fonds publics sont en jeu, sont très importantes, et les chercheurs, le personnel de soutien, les organismes de recherche et leur agence de financement doivent alors établir des rapports complexes. Au début du rapport, on examine la situation et on présente une démarche modeste qui protège le savoir collectif ainsi que les données expérimentales et qui répond aux besoins de chaque groupe. Cette démarche a pour point de départ la codification des métadonnées tirées des expériences et leur représentation en XML. On a recours au langage XML, car il est reconnu comme le mécanisme standard de représentation, de transformation et de partage de l'information descriptive informatisée.

La démarche est mise en pratique dans une application appelée *PI Assist*. Dans le rapport, on donne un bref aperçu du système, puis on discute des résultats du premier essai sur place de ce dernier. Il a ainsi été possible de cerner un certain nombre d'améliorations à apporter. Celles-ci ont été intégrées à la version actuelle qui peut maintenant être soumise au second essai sur place.

Dans le cadre de ces travaux, NTT s'est également penché sur des démarches visant à rechercher des collections de documents dans *PI Assist*. Une technique faisant appel au système de gestion de bases de données ORACLE a alors été conçue et mise en place. Dans l'ensemble, les résultats n'ont pas été satisfaisants, surtout pour des raisons concernant l'immaturité relative de la technologie de combinaison de la base de données et du langage XML. Comme il n'existe aucun besoin actuel lié à cette installation de recherche, étant donné le faible nombre de documents disponibles, nous recommandons de suspendre les travaux dans ce secteur, et ce jusqu'à ce que les solutions des tierces parties évoluent davantage.

Le rapport fait état finalement de la nécessité de soumettre l'application à des essais auprès des scientifiques et recommande que le personnel de soutien d'ECC entreprenne d'utiliser *PI Assist* pour documenter les expériences déjà menées.

Executive summary

This report describes the status of *PI Assist*, a knowledge management tool developed for the Command Effectiveness and Behaviour Section at DRDC Toronto. The system is designed to help researchers create and document their experiments, support-staff plan for resources and configure equipment, and analysts locate and process data and document their results.

In doing this, *PI Assist* helps individuals and teams conduct their research while at the same time accruing significant benefits to research organizations. These arise in part from capturing the knowledge created and used in these activities, and in part from the protection this in turn provides for the data asset created over the course of the studies.

The report describes the rationale that underlies *PI Assist* and briefly describes the system. It then discusses the system's current status, including the results of a preliminary field trial, and enhancements made in its wake.

The report concludes by recommending immediate use of the system for the purposes of documenting and archiving completed experiments. It further recommends a second field trial with scientists to further evaluate *PI Assist*'s usefulness as an aid to creating and documenting experimental designs.

Grushcow, M.S. 2004. CEB Section Experiment Archiving Tool. CR 2004-069 DRDC Toronto.

Sommaire

Le présent rapport décrit l'état de *PI Assist*, un outil de gestion du savoir qui a été développé à l'intention de la Section de l'efficacité du commandement et du comportement à RDDC Toronto. Le système vise à aider les chercheurs à créer et à documenter leurs expériences, le personnel de soutien à planifier les ressources et à configurer l'équipement, et les analystes à trouver et à traiter les données ainsi qu'à documenter leurs résultats.

Ainsi, *PI Assist* aide les personnes et les équipes à mener leurs travaux de recherche, tout en offrant encore plus d'avantages importants aux organismes de recherche. Ces avantages découlent, d'une part, de la saisie du savoir créé et utilisé dans ces activités, et, d'autre part, de la protection alors ainsi assurée pour l'ensemble des données créé au fil des études.

Le rapport décrit les motifs ayant donné lieu à *PI Assist* et présente brièvement le système. Il traite ensuite de l'état actuel du système, y compris les résultats d'un essai sur place préliminaire, de même que les améliorations qui y ont été apportées dès les premiers débuts.

Le rapport recommande finalement que le système serve immédiatement à documenter et à archiver les expériences menées. Il recommande aussi qu'il soit soumis à un second essai auprès des scientifiques, afin d'évaluer davantage *PI Assist* et son utilité comme aide à la création et à la documentation de concepts expérimentaux.

Grushcow, M.S. 2004. CEB Section Experiment Archiving Tool. CR 2004-069 DRDC Toronto.

Table of contents

Abstract i

Résumé ii

Executive summary iii

Sommaire iv

Table of contents v

Introduction 1

- The Problem 1
- The Researcher's View 1
- The Organization's View 3
- The Funder's View 4
- Synthesis 4
- A Modest Approach 5
- The Solution 5
- Benefits 6

PI Assist – The Program 7

Current Status 8

Conclusion and Recommendations 10

Introduction

The Problem

PI Assist is a document-based Archiving and Knowledge Management System for scientific research studies. It allows for the organization and inter-relation of all the information that goes into and comes out of a research project, all in a single document.

PI Assist was developed as a way of dealing with CEB's data archiving needs. The section has been responsible for a number of large and small, related experimental studies that date to the late 1970's. Data was acquired and stored in a way that allowed results from all studies over the 25-year period to be read, combined and manipulated.

Unfortunately, particularly for the early experiments, we could not always identify sufficient experimental design information to be able to interpret the data.

In other words: *Our ability to interpret, share, and reuse data depends on our having knowledge of the context in which it was collected.*

Since we could no longer specify the design and manipulations that were in effect when the data were collected, we couldn't use them for comparative and retrospective studies. The *data's* value had depreciated to near zero.

This probably isn't an unusual situation in labs around the world. Different researchers document their work to different levels of detail and completeness. The reason we have better success interpreting more recent studies is not because researchers are more careful about documenting their work now than they used to be. It is because many of those who worked on the recent studies, particularly support staff, are still with the organization. They can remember the details of the studies, or know where to find the notebooks, and so on. This brings us to a realization:

Data persists, Knowledge decays.

But how does this affect the various parties?

The Researcher's View

When we are motivated to revisit an experiment, we are immediately faced with the issue of locating and recovering a mass of contextual information not recorded in the publications, such as:

- the references that informed the work.
- the overall design (and possibly earlier abandoned approaches), including factors and hypotheses.
- how the data was manipulated prior to detailed statistical analyses including methods for handling missing values.

- anomalies that occurred during the experiment that may have affected the data.
- information required to replicate and defend the statistical results.

Since designing and running an experiment are highly collaborative activities, our ability to answer these questions depends on information provided by several groups, including collaborators and research staff. How does the Principal Investigator (PI) ensure that this is provided, stored and retained?

One might ask *why* the PI cares. There are a number of obvious reasons including professionalism, the need to defend results during peer reviews, and lab policies. These all fall into the class of things that *ought* to be done—that is, obligations to others and to the institutions for which we work.

However, there are many other reasons to care that are nearer and dearer to the scientific heart of a researcher. Efficient organization and retrieval of information will make the PI much more personally productive. Controlling the data and being able to track how it was processed and analyzed simply makes answering questions easier and faster. Retention of early design decisions, literature search results, and the overall and detailed final design all support the time efficient development of follow-on research. This is all the more true the longer the gap between the original and new work.

For the researcher, information about past experiments is necessary since it supports new research. It does this by providing useful data and analyses, and more broadly, experimental designs, hypotheses, theories, and conclusions. Note that the usual source for all but the first is normally personal notes, collective memory, and the published scientific literature.

Even when follow-up work is being done by the same researcher, it can save much time and re-treading of old ground to have the experimental design, its connections to the hypotheses and the actual data and conclusion all there in one place, along with notes about future work—all ready to be examined, manipulated or re-worked. This is particularly true if the researcher is reviving an old project that was completed ten years ago. And, of course, this is all even more pertinent if the researcher is doing a follow-up study of another PI's experiment.

In addition to the benefits of simply having one's contextual information ready-at-hand, there are further benefits when this information is pre-structured in a way that reflects scientific procedures. Such pre-structuring makes entering the data a natural act, and provides the pleasant side-benefit of helping to train new support staff and assistant researchers on the essentials of conducting good research, since they have a structured template for the kind of information they are expected to collect and maintain. There is a placeholder already there for the experimental design elements, for hypotheses, for references, for independent variables and dependent measures, and so on. Since all this is structured to reflect how a scientist works, the resulting document is not just an efficient information manager, but an integral part of the entire process of conceiving, designing and carrying out the study.

The Organization's View

Some of the PI's concerns are directly shared by the organization for which he or she works. For example, if data retention is mandated, a lab may be obliged to put appropriate procedures for this in place. However, the auditing requirement will usually be for the ability to physically or electronically deliver the data on demand; not necessarily to be able to interpret it in a way that supports current and future research. Auditing rules are rarely, if ever, designed to benefit an organization's ongoing business. Similarly, since a research organization's reputation is based on the work of its researchers, labs frequently put procedures in place that assist the analytical and publishing processes. But again, these deal with the last experiment, not the next one.

In terms of ongoing work, an organization is concerned with funding the activities of a number of researchers, many of whom share lab facilities, staff, and other resources. The more efficiently it can do this, the better. This means establishing an infrastructure that can provide services ranging from finding suitable subjects, to configuring computers and data acquisition systems, overseeing experimental sessions, collecting and cleaning raw experimental data, and so on. Typically, research assistants and related support staff do most of this work. Frequently, they design the procedures themselves and effectively create and manage the organization's "experiment running business." Efficiency is a measure of how well they organize and systematize their work. These procedures evolve over time as a result of experience and changing experimental needs, representing a significant body of knowledge that is often systematized in very informal ways including documented and undocumented procedures, mentoring, so on.

Research assistants learn from the last experiment and apply their experience to the next. Like the researchers themselves, their focus is forward. They can tell you what their data cleaning procedures are now, and likely have that information documented. But the documentation itself has evolved, and finding out what the procedures were four experiments ago might be problematic. This has serious implications for data reusability.

While research managers want to provide good research environments for ongoing work, they also have broader organizational concerns. What are the products of the research? They clearly go beyond the publications. Where is the record of all the science that never made it to publication? When researchers collaborate on an experiment, the overall design embodies all of their hypotheses. When the work is completed, a number of papers are published, each dealing with a specific aspect of the experiment. What captures the knowledge about the experiment in its totality? This includes the things that didn't work, as well as the serendipitous findings.

For research managers, information about past experiments including the data, their design, the history of their execution, and so on, all represent assets that the organization builds up over time. There is a huge investment in knowledge, both scientific and operational. How is this knowledge to be organized, protected, managed, and made available to support future work? What keeps it from "depreciating?" How does it survive personnel transitions?

The Funder's View

Funding agencies share many concerns with the organizations that they fund. However, their focus is less on the operational side, their time scale longer, and their viewpoint crosses organizational boundaries. Historically, although they fund the research, they receive only a part of the knowledge that is created. This is typically represented by the "final report" that presents the findings. The main impact of this is that although they fund similar work in multiple organizations, there is no mechanism for each project to benefit from the detailed knowledge that each of the others generates. This effectively wastes resources due to replication, loss of synergy, and the inability to identify best practices.

For the funding agency, information about past experiments (not just results) represents knowledge that is paid for but stays with the receiving organization. This is particularly an issue where public funds are involved. How can the information be systematized and shared?

Note that in the context of DRDC, some of these issues, particularly the need for agency wide sharing, are mitigated by the fact that the different labs represent different areas of interest. Consequently the value of results sharing at this level is not as great as it would be in areas where different groups conduct similar research. Nonetheless, issues relating to managing the results of publicly funded research remain.

Synthesis

None of the above is meant to imply that researchers aren't concerned about the scientific or financial impact of how they do their work; or, that their organizations are only concerned with administrative procedures; or, that funders are only interested in costs and deliverables. Each of these actors is operating in their own time scale and is driven by their own priorities.

Everyone would benefit if we could have a collection of experiment descriptions, each of which:

- identified the individuals and organizations involved in the work.
- contained an overview of the experimental goals and the references that informed the new work.
- provided a detailed description of hypotheses, factors and dependent variables.
- described the subject pool in terms of relevant demographics.
- specified the experimental protocol, facilities used, task descriptions, and run histories.
- identified and described the data acquisition environment.
- allowed direct access to a variety of supporting files including raw and cleaned data, statistical summaries, and related electronic documents.
- contained full descriptions of data cleaning procedures.
- included annotated analytical results.
- gave conclusions, dealing with each of the hypotheses and the experiment as a whole.
- could be represented in a format that maximized portability and information sharing.

- could be created without adding work at any level that did not have real benefit.

Investigators would benefit from a wealth of richly documented data for retrospective analysis, and detailed experimental designs that could inform new studies.

Operational and procedural information, collected in the course of configuring, running and analyzing experiments, would meet the organization's need for capturing the knowledge that describes how the work was done.

On a more global level, the experiment descriptions in their entirety would fully document the studies conducted—something a final paper can never do—and would truly embody the organization's intellectual assets. The contextual information that they contain would provide the structure and detail needed to interpret the underlying data. Funding agencies would receive not only broad descriptions and conclusions, but also detailed, structured, and portable information that could be shared by their research organizations.

A Modest Approach

The situation we have described is commonly referred to as a *Knowledge Management* problem (note the capitals). However, if we limit our goals, we can provide a reasonable solution at a reasonable cost, with significant benefits to researchers, their organizations and their funding agencies.

Our goal in designing the current version of *PI Assist* was to produce an *archive* for a study that:

1) describes the experiment including
    a. goals and design,
    b. resources employed including their configuration,
    c. data collected along with a description of their formats,
    d. data cleaning and analysis procedures,
    e. results,
    f. conclusions.
2) is portable and easily accessible through Internet technologies.
3) can be interpreted using standard tools.
4) is produced over the course of creating, running and interpreting the experiment.
5) should not require significant extra effort on the part of those involved.
6) provides significant benefits for those who create the archive, as well as others.

The Solution

To reach these goals, we realized that we had to focus less on the data itself, which is easy to organize, and more on the metadata—information about the data. This meant designing a rich and highly structured document template for the metadata, one that reflects the way researchers actually do science. In a sense, the scientific method itself would need to be

embedded in the structure of the template, to guide authors through the document creation process, and make it part and parcel of the process of doing their work.

To maximize support for sharing experiment descriptions between collaborators and institutions, we decided to use XML, the industry standard for structured documents and data sharing.

While the document template reflects the *process* of doing science, the final paper represents the *communication* of the result. While this communicated result may have a somewhat different form than the study itself, the connection between the two is still important. In addition, while each individual researcher may understand the information they are entering at the time they enter it, the entire sum of all the information in the document still needs to be communicated in a readable way, both to external researchers and to the various individuals, each working on their own piece of the same study. To this end, we recognized the need for an automatically generated manuscript-like view of the structured metadata. This would provide, not a final paper, but a way of communicating the information in the document *in a readable way*, and in a way that might provide a starting point for the creation of a final paper.

In addition, we realized that by combining experiment description with data organization, within a single document, a research project gets archiving and document version control "for free". At certain intervals throughout a project, the ongoing study documents can be archived away to storage, and at the end of the project, we have not only a complete record of the study, but a record of what stage it was in throughout its life cycle, from conception to publication.

Benefits

The *PI Assist* document template solicits and organizes critical operational and scientific knowledge, thus capturing context, intent, process and results.

Because there will always be information that it is not practical to physically store in the study document, *PI Assist* has the ability to reference external data files, papers, and so on, so they may all be retrieved and viewed from the study document, either directly within *PI Assist* or in a standard Web browser.

Both data and metadata can be easily transferred to archival storage.

Because the document is in XML, any number of reports could in theory be created, and different information could be extracted for different audiences. The information could be reorganized and reformatted for different media and publication styles. While there is currently no built-in support for this level of customizability, the use of XML ensures that this feature can be added in a future version of the software with little change to what is already there. Thus XML gives the software the room it needs to grow in capability and flexibility in the future.

PI Assist – The Program

PI Assist presents users with a single interface that combines document structure, customized editors, and a view of the complete document.

The main *PI Assist* window (Figure 1) has three areas, two of which are Views of the Study document (one for navigating and the other for presentation), and the other is an editor for entering and editing the data contained in the document:

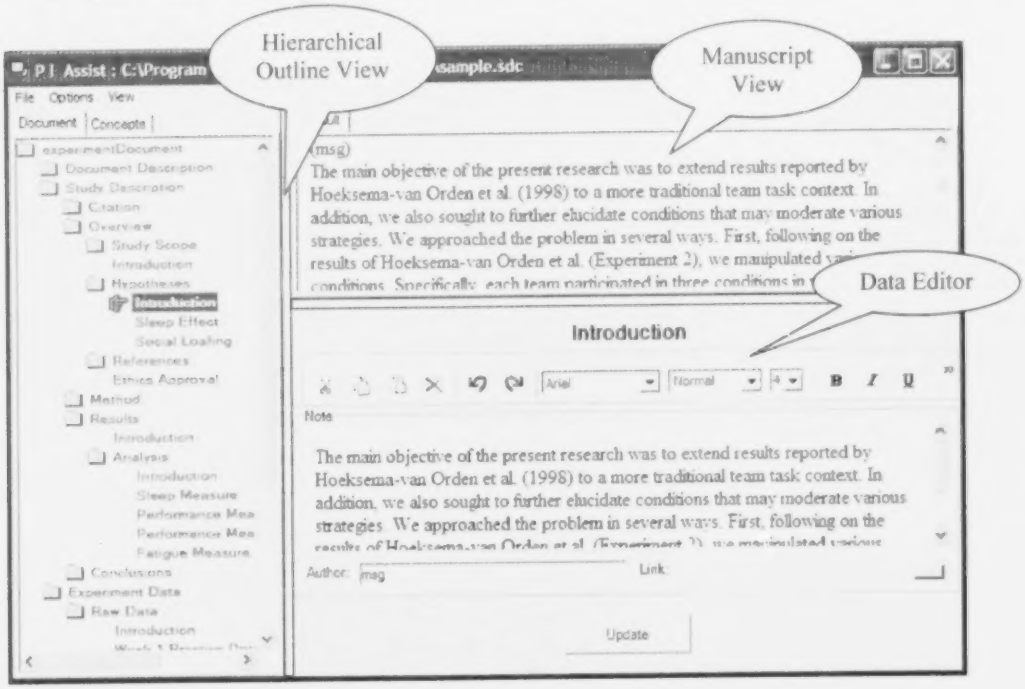
Hierarchical Outline View
Manuscript View
Data Editor
P.I. Assist : C:\Program
sample.sdc
File Options View
Document Concepts
experimentDocument
Document Description
Study Description
Citation
Overview
Study Scope
Introduction
Hypotheses
Introduction
Sleep Effect
Social Loafing
References
Ethics Approval
Method
Results
Introduction
Analysis
Introduction
Sleep Measure
Performance Mea
Performance Mea
Fatigue Measure
Conclusions
Experiment Data
Raw Data
Introduction
(msg)
The main objective of the present research was to extend results reported by Hoeksema-van Orden et al. (1998) to a more traditional team task context. In addition, we also sought to further elucidate conditions that may moderate various strategies. We approached the problem in several ways. First, following on the results of Hoeksema-van Orden et al. (Experiment 2), we manipulated
Introduction
Arial
Normal
Note
The main objective of the present research was to extend results reported by Hoeksema-van Orden et al. (1998) to a more traditional team task context. In addition, we also sought to further elucidate conditions that may moderate various strategies. We approached the problem in several ways. First, following on the
Author: msg
Link
Update

Figure 1. *The Main Window: Hierarchy/Manuscript Views and Data Editor.*

Outline View: Positioned at the left of the main window, this is a standard hierarchical outline view of the document's structure, consisting of main headings, subheading, and so on. Each heading can be expanded to view its internal structure, or collapsed to hide it. Navigating down into this hierarchy, leads to headings that do not expand—these are Data Headings, which contain the actual data. None of the actual data is presented for viewing in the outline view, however—only the headings appear here.

Manuscript View: This view (in the upper-right of the window) presents the information contained in the document in a readable manuscript form. It is not intended to be a final manuscript; rather it is simply a view of the study presented in roughly that form. It may, however, be saved to HTML and opened in a word processor to serve as the starting point for a final paper.

Data Editor: Whenever a Data Heading is selected in the Outline View, its corresponding editor appears in the lower right of the main window. There are many different data editors, each one tailored to the kind of data expected at the corresponding heading.

Concept View: This is an alternative to the Hierarchy View, also displayed on the left-hand side of the window. Choose between the two by clicking on the corresponding tabs at the top of the View. This View relates to the Study Document and the other Views in much the same way as the Hierarchy View, except that it is flat (non-hierarchical) and organized alphabetically. It thus functions has a handy index.

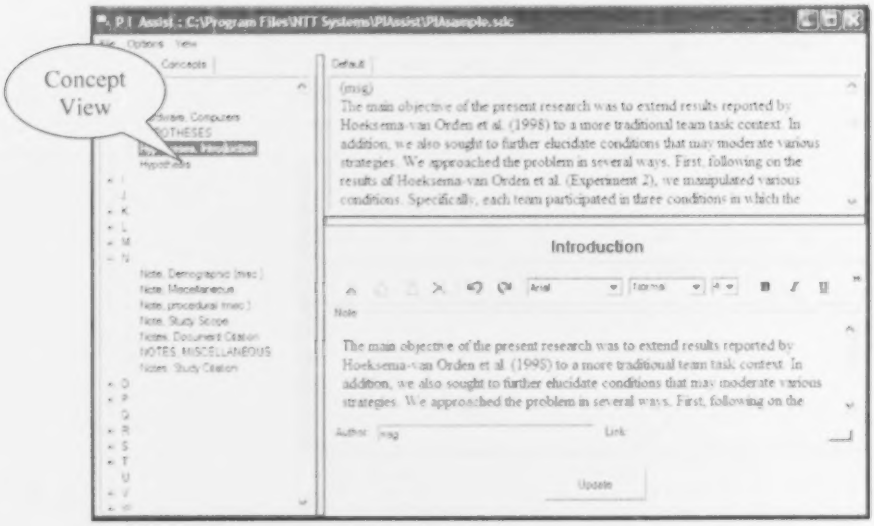

Figure 2. *The Concept View.*

A complete overview, user manual and context sensitive help system are separately available.

Current Status

PI Assist was released for a limited trial within the Section in September 2003[1]. Results are available in a separate contractor's report[2]. The initial response showed more acceptance of the system as an archiving tool than as an experiment design tool. The trial evaluation identified a number of usability improvements that addressed these shortcomings. These included: the development of an extensive online help system, the provision of concept based navigation as an alternative to the document hierarchy, and a feature that allows for the

[1] W7711-7687(20)

[2] Grushcow, Marc S.. 2003. Initial Experiences Using PI Assist in the Command Effectiveness & Behaviour Section. DRDC - Toronto

creation of new attached documents directly inside the archive. The new system will be released for another trial in April 2004.

Part of the work called for under this contract related to the development of a database that would house *PI Assist* documents in a way that supported "intelligent searching." Intelligent searching is meant to convey the idea that the search has to consider both text and context. For example, we want to find studies where Baranski is the Principal Investigator, and not ones in which he appears as an author on a referenced paper. *Google* type searches based on html representations of the document obviously cannot make this kind of distinction. However, since the document is represented in XML, the association between text and context is clear. The problem is that there are no established search mechanisms that actually provide this capability for files stored in directories[3].

We investigated both Microsoft SQLServer and the Oracle database management system to see if they could provide a solution. SQLServer was eliminated based on certain technical issues[4]. We did successfully store complex sample documents in an Oracle database. This took considerable work involving the most current release of Oracle 9, and related patches and updates.

We then created a search interface within *PI Assist* itself. This allows users to specify a variety of criteria and initiate a search. Titles from matching documents are presented for review and these can be read from the database and opened for review inside *PI Assist*. A view of the search interface is shown below:

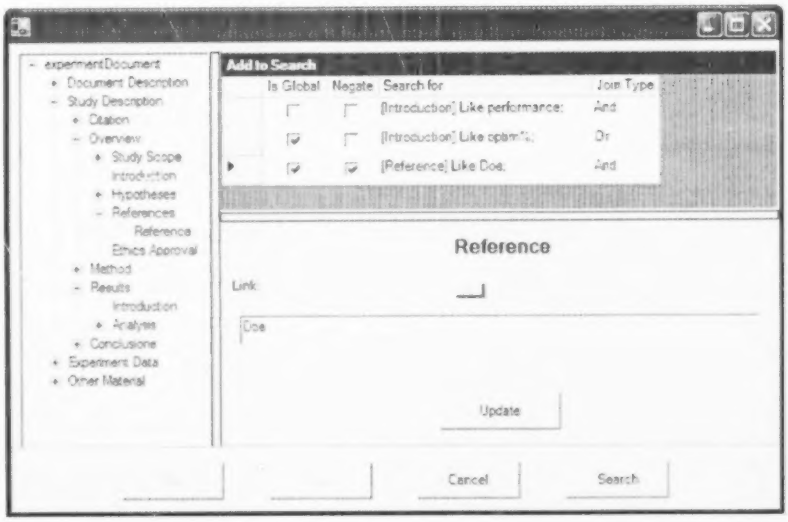

Figure 3. *A Compound Search.*

[3] Notwithstanding some of the advertising claims.
[4] Primarily its inability to store mixed content inner text fields.

The search facility works and can be demonstrated. However, due to the current complexities associated with inserting information into the database, and the logistics associated with supporting it, we do no recommend proceeding further at this time. This has no operational impact since the Section currently has no completed *PI Assist* documents. We have demonstrated the capability and improved vendor support will make the system more usable by the time it is needed.

Conclusion and Recommendations

1. Archiving and Experiment Design

   Recent experiences in the Section reinforce the need for *PI Assist*, at least as an archiving tool for completed experiments. This work should proceed as soon as possible using the new software release. Another trial should be conducted with scientists to determine the extent to which the improvements meet their needs. If scientist can effectively use the tool during the process of designing experiments, the overall quality and completeness of the documents will be much better than it would be if support staff had to do this work after the fact.

2. Database Archiving and Intelligent Search Engines

   We recommend delaying further work on this feature due to the state of the technology. Technically, the capability exists to do this, but it is in its infancy. The growing use of XML as an Internet document format will increase the pressure for the development of intelligent search mechanisms. We may see this develop using either or both of file and database repositories. In the absence of a pressing need, waiting for the technology to mature will be both cost and effort efficient.

3. Pre and Post Experiment Activities

   The SUSOPS system provides a mechanism for distributing and installing the files needed to run a particular study. It does not address the general problem of dealing with other types of experiments (e.g., TITAN). Moreover, we have never addressed the question of collecting and packaging data from the set of computers used in any given study. These activities are obviously related to *PI Assist* and, more broadly, distribution and archiving.

   Post experiment activities, i.e., collecting, packaging, and annotating raw data, can and should be integrated into *PI Assist* as soon as possible. This will greatly assist support staff that will be retrospectively documenting old experiments.

   Pre experiment activities, i.e., assembling and installing the files and programs needed to run a study, is more complex. This is in part due to the different dependencies and requirements of the individual test tasks. One approach to this problem would be to find, or create, a general packaging system that could use templates for the different types of studies we run. This would give us a single distribution platform that could effectively be customized for different types of

experiments. Pursuing this path should follow a review within the Section of research activities proposed for the next few years.

In summary, we do not recommend any further activity relating to using databases to store *PI Assist* documents. Too much effort would be needed at this time, and any local solution would be overtaken by open source and commercial offerings by the time the Section actually needed it.

The general question of how experiments are distributed and their data collected should be investigated after the Section reviews its mid-range research plans. The prime benefit of having a packaging and distribution system would be in increasing the support staff's ability to independently manage this process.

Finally, we recommend using *PI Assist* within the support section for documenting completed studies. An additional enhancement to facilitate the collection and packaging of raw data should be undertaken. A further field trial of *PI Assist* as an experiment design tool should also be undertaken. This will determine the extent to which the new system addresses concerns raised in the last trial.

DOCUMENT CONTROL DATA SHEET

1a. PERFORMING AGENCY
NTT Systems Inc., 86 Dunblaine Ave., North York, Ontario., M5M 2S1

2. SECURITY CLASSIFICATION
UNCLASSIFIED

1b. PUBLISHING AGENCY
DRDC Toronto

3. TITLE
(U) CEB Section Experiment Archiving Tool - A Knowledge Management Approach to Designing and Archiving Experiments and Their Data

4. AUTHORS
Marc S. Grushcow

5. DATE OF PUBLICATION
March 31 , 2004

6. NO. OF PAGES
23

7. DESCRIPTIVE NOTES

8. SPONSORING/MONITORING/CONTRACTING/TASKING AGENCY
Sponsoring Agency:
Monitoring Agency:
Contracting Agency : DRDC Toronto
Tasking Agency:

9. ORIGINATORS DOCUMENT NO.
Contract Report CR 2004-069

10. CONTRACT GRANT AND/OR PROJECT NO.
W7711-0- 7687

11. OTHER DOCUMENT NOS.

12. DOCUMENT RELEASABILITY
Unlimited distribution

13. DOCUMENT ANNOUNCEMENT
Unlimited announcement

14. ABSTRACT

(U) The Command Effectiveness and Behaviour Section at DRDC-Toronto has been conducting cognitive psychology and team related experiments since the late 1970's. Over that time it has amassed a large volume of raw and analyzed data from a series of related, large and small research studies. Technically, data from different studies can be combined for analysis because they are all sufficiently well identified with respect to experiment, subject, task, trial, etc. However, this has not turned out to be the case. The reason is that, particularly for the early experiments, there is no longer sufficient documentation, or corporate knowledge, to reconstruct the experimental designs. This absence of knowledge of the context in which the data were collected renders them useless for the purpose of further analysis.
The reliance on staff as the repository of corporate knowledge is not unusual. But the consequences in research environments, particularly where public funding is involved, are far reaching and involve complex relationships between researchers, support staff, research organizations, and their funding agencies. The report begins by reviewing this situation and lays out a modest approach that protects corporate knowledge and experimental data while addressing the needs of each group. The central features of the approach are codification of experiment metadata and their representation in XML. XML is used since it is recognized as the standard mechanism for representing, transforming, and sharing computer-based descriptive information.
The approach is implemented in an application called PI Assist. The report presents a brief overview of the system and then discusses the results of its first field trial. This identified a number of enhancements that were implemented for the current version that is now ready for the second field trial.
In the course of pursuing this work, NTT also investigated approaches to searching collections of PI Assist documents. An implementation using the ORACLE database management system was designed and implemented. This was not altogether satisfactory, primarily for reasons relating to the relative immaturity of the combined database/XML technology. Since there is no current need for this search facility, owning to the small number of documents available, we recommend suspending work in this area until third party solutions evolve further.
The report concludes by identifying the need for further field trials with scientists and recommending that support staff in CEB begin using PI Assist to document experiments that have already been completed.

(U) La Section de l'efficacité du commandement et du comportement à RDDC Toronto mène depuis la fin des années 1970 des expériences liées aux équipes et à la psychologie cognitive. Pendant cette période, elle a accumulé un volume important de données brutes et analysées à l'aide d'une série d'études de recherche connexes, de grande et de faible envergures. Sur le plan technique, il est possible de combiner les données provenant d'études différentes, car l'expérience, l'objet, la tâche, les essais, etc. de celles-ci sont assez bien définies. Par contre, cela n'a pas été possible ici. En effet, en ce qui concerne plus particulièrement les premières expériences, il n'y a plus assez de documentation, ou de savoir collectif, pour reconstruire les concepts expérimentaux. Étant donné un tel manque de contexte quant à la façon dont les données ont été recueillies, il est donc inutile d'analyser davantage celles-ci.
Il n'est pas rare de recourir et se fier au personnel pour qu'il joue le rôle de dépôt du savoir collectif. Toutefois, les conséquences d'une telle décision dans des environnements de recherche, en particulier lorsque des fonds publics sont en jeu, sont très importantes, et les chercheurs, le personnel de soutien, les organismes de recherche et leur agence de financement doivent alors établir des rapports complexes. Au début du rapport, on examine la situation et on présente une démarche modeste qui protège le savoir collectif ainsi que les données expérimentales et qui répond aux besoins de chaque groupe. Cette démarche a pour point de départ la codification des métadonnées tirées des expériences et leur représentation en XML. On a recours au langage XML, car il est reconnu comme le mécanisme standard de représentation, de transformation et de partage de l'information descriptive informatisée.
La démarche est mise en pratique dans une application appelée PI Assist. Dans le rapport, on donne un bref aperçu du système, puis on discute des résultats du premier essai sur place de ce dernier. Il a ainsi été possible de cerner un certain nombre d'améliorations à apporter. Celles-ci ont été intégrées à la version actuelle qui peut maintenant être soumise au second essai sur place.
Dans le cadre de ces travaux, NTT s'est également penché sur des démarches visant à rechercher des collections de documents dans PI Assist. Une technique faisant appel au système de gestion de bases de données ORACLE a alors été conçue et mise en place. Dans l'ensemble, les résultats n'ont pas été satisfaisants, surtout pour des raisons concernant l'immaturité relative de la technologie de combinaison

de la base de données et du langage XML. Comme il n'existe aucun besoin actuel lié à cette installation de recherche, étant donné le faible nombre de documents disponibles, nous recommandons de suspendre les travaux dans ce secteur, et ce jusqu'à ce que les solutions des tierces parties évoluent davantage. Le rapport fait état finalement de la nécessité de soumettre l'application à des essais auprès des scientifiques et recommande que le personnel de soutien d'ECC entreprenne d'utiliser PI Assist pour documenter les expériences déjà menées.

15. KEYWORDS, DESCRIPTORS or IDENTIFIERS

(U) archive; knowledge management